학부졸업논문

데이터 불균형 문제 해결을 위한 오버샘플링 방법론에 관한 연구

A study on Oversampling Methodology for Resolving Class Imbalanced Problems

전남대학교 공과대학

산 업 공 학 과

문 혁 , 임 창 엽

지도교수 정 영 선

2026년 06월

데이터 불균형 문제 해결을 위한 오버샘플링 방법론에 관한 연구

A study on Oversampling Methodology for Resolving Class Imbalanced Problems

이 논문을 공학 학사학위 논문으로 제출함

전남대학교 공과대학

산 업 공 학 과

문 혁 , 임 창 엽

지도교수 정 영 선

문혁과 임창엽의 공학 학사 학위논문을 인준함

심사위원장 정 영 선 (인)

2026년 06월

목차

표 목차

그림 목차

데이터 불균형 문제 해결을 위한 오버샘플링 방법론에 관한 연구

A study on Oversampling Methodology for Resolving Class Imbalanced Problems

문 혁 , 임 창 엽

전남대학교 산업공학과
(지도교수: 정 영 선)

(국문초록)
본 연구는 분류 문제에서의 데이터 불균형을 해소하기 위해, 마할라노비스 거리와 고유 선분 중심 게이팅 함수를 결합한 오버샘플링 방법론을 제안한다. 기존 선형 보간 기법들이 데이터의 고유 공분산 구조를 반영하지 못하고 다수 클래스 영토를 침범하는 클래스 중첩문제를 일으키는 한계를 극복하고자, 본 연구에서는 Fast-MCD 알고리즘을 통해 이상치가 배제된 마할라노비스 거리 공간을 구축한다. 이후 소수 클래스 간의 고유 선분을 구성하고, 각 선분의 중점과 분포 중심 간 마할라노비스 거리의 역수를 생성 가중치로 정의한다. 이를 통해 확률 밀도가 높은 코어 영역의 생성량은 극대화하고, 클래스 중첩 위험이 높은 외곽 선분의 생성량은 최소화하는 오버샘플링을 수행한다. KEEL 저장소의 불균형 벤치마크 데이터 세트 6종을 대상으로 SVM 분류 모델을 적용해 실험한 결과, 제안 기법은 대조군 대비 F1-Score의 향상을 달성하였으며, 유의수준 0.05 하에서 통계적 유의성을 검정하였다.

주요어: 불균형 데이터, 데이터 오버샘플링, 마할라노비스 거리, 최소공분산결정, 클래스 중첩

1. 서론

1.1. 연구 배경

최근 다양한 산업 분야에서 머신러닝 기반 예측 모델이 증가하고 있다. 그러나 실제 산업 현장에서 수집되는 데이터는 양성 클래스의 비중이 음성 클래스 대비 극히 적은 데이터 불균형(Data Imbalanced) 특성을 보인다. (Vinoodhini, D., 2022). 예를 들어 제조 공정에서는 생산 제품이 대부분 양품으로 구성되며 불량품은 극히 적은 비율만 존재한다.

이와 같은 데이터 불균형 문제는 머신러닝 모델의 학습 과정에서 다수 클래스에 대한 편향을 유발한다. 대부분의 분류 알고리즘은 전체 오분류율을 최소화하는 방향으로 학습이 이루어지므로, 다수 클래스에 편향되어 소수 클래스에 대한 식별 능력이 저하된다 (Krishna & Sidharth, 2022).

특히 데이터 불균형 환경에서는 정확도(Accuracy)만으로 모델의 성능을 평가할 경우 정확도의 역설이 발생할 수 있다. 예를 들어 전체 데이터의 90%가 다수 클래스인 경우 모든 데이터를 다수 클래스로 예측하더라도 90%의 높은 정확도를 얻을 수 있다. 즉 전체 정확도는 높게 나타나지만 2종 오류(False Negative; FN)를 탐지하는 Recall 수치가 현저하게 떨어져 실전 예측 도구로서의 신뢰성이 극히 낮아지게 된다 (Krishna & Sidharth, 2022).

1.2. 연구 목적 및 필요성

데이터 불균형 문제를 해결하기 위한 대표적인 방법으로 SMOTE를 비롯한 다양한 기법들이 제안되어 왔다. 이후 데이터의 분포 구조를 반영하기 위해 마할라노비스 거리를 활용한 오버샘플링 기법이 연구되었으며, 최근에는 이상치에 의한 공분산 행렬 왜곡 문제를 완화하기 위해 MCD를 결합한 방법론도 제안되었다.

그러나 이러한 방법론들도 여전히 선형 보간 기반 데이터 생성으로 인해 다수 클래스 영역을 침범하는 클래스 중첩(Class Overlap) 문제를 충분히 해결하지

못하였다. 따라서 본 연구에서는 MCD 기반 마할라노비스 거리 공간을 활용하면서도, 각 고유 선분의 위치적 특성에 따라 생성되는 인공 데이터의 양을 차등적으로 조절하는 새로운 오버샘플링 방법을 제안하고자 한다.

1.3. 연구 구성

본 논문은 다음과 같이 구성되어 있다. 2장에서는 관련된 선행연구를 고찰하며 본 연구의 차별성을 파악한다. 3장에서는 데이터 불균형 문제에 대한 내용, 데이터 불균형 문제 해결을 위한 데이터 증강 기법 그리고 마할라노비스 거리에 대한 이론적인 배경을 서술한다. 4장에서는 본 연구에서 제안하는 오버샘플링 기법을 설명한 후, 5장에서는 실험에 사용된 데이터의 소개, 기존의 방법론과의 성능 비교를 제시하고 마지막으로 통계적 유의성 검정을 실시한다. 마지막으로 6장에서는 본 연구에 대한 결론과 한계점, 향후 연구 방향에 관해 서술한다.

2. 선행연구 고찰

2.1. 선행연구 동향

데이터 불균형 문제를 해결하기 위한 데이터 수준 접근법에 관한 연구는 국내와 해외 모두 꾸준히 진행되고 있다.

Chawla et al.(2002)는 대표적인 데이터 수준 접근법 중 하나인 SMOTE(Synthetic Minority Over-sampling Technique)를 제안하였다. SMOTE는 소수 클래스 내 임의의 두 샘플 간의 선형 보간을 통해 가상의 샘플을 생성하는 오버샘플링 기법이다. 이는 단순 무작위 오버샘플링이 초래하는 과적합 문제를 해결하고 국지적 밀도 영역 확장을 통한 소수 클래스 경계면을 공고히 하였으나, 소수 클래스의 공분산 구조를 반영하지 않기 때문에 분포 특성을 반영하지 못하고 다수 클래스 영토를 침범하는 클래스 중첩 문제를 유발한다 (Zhang et al., 2023).

Bennin et al.(2018)는 합성 표본의 다양성을 증대시키기 위해 복수 부모로부터 특징을 재조합하는 유전학과 유사한 전략을 도입하여 SMOTE 계열에서 발생하는 중복, 과집중 문제를 완화하고 소수 클래스의 표현 범위를 넓히는 'MAHAKIL'을 제안하였다.

Abdi와 Hashemi(2016)는 기존의 SMOTE 기법에 마할라노비스 거리를 적용한 MDO(Mahalanobis Distance based Oversampling) 알고리즘을 제안했다. 이 알고리즘은 유클리드 거리를 기준으로 K개의 최근접 이웃 데이터를 선택하는 방법 대신 전체 데이터 세트의 공분산을 반영한 마할라노비스 거리를 기준으로 K개의 최근접 이웃 데이터를 선택하였다. 하지만 소수 클래스 내 공간적 이상치, 노이즈가 존재할 경우 데이터의 공분산 구조가 왜곡될 수 있는 문제가 존재한다.

이상치에 의해 공분산 구조가 왜곡되는 것을 방지하기 위해 정지은과 최용석(2024)는 Fast-MCD(Minimum Covariance Determinant)를 이용하여 평균벡터와 공분산 행렬을 추정하고 이를 마할라노비스 거리 계산에 적용한 MCD-SMOTE를 제안하였다. 이는 이상치에 의한 평균과 공분산 행렬의 왜곡을

최소화할 수 있고, 왜곡이 최소화된 마할라노비스 거리를 활용한 MDO를 수행할 수 있게 한다. 그러나 정렬된 선분 위에서 수행되는 SMOTE로 인하여 클래스 중첩(Class Overlap) 문제는 여전히 해소하지 못하였다.

2.2. 연구의 차별성

기존의 선행연구들은 데이터의 고유한 분포 구조를 반영하기 위해 마할라노비스 거리를 사용하였으며, 이상치로 인한 공분산 행렬 왜곡 문제를 해소하기 위해 최소공분산결정(MCD) 추정량을 결합하였다. 이를 통해 구성된 마할라노비스 거리 공간 내에서 최근접 이웃(K-NN) 간의 선형 보간을 통한 오버샘플링을 수행하였다.

그러나 선행연구들이 데이터의 분포 구조 반영과 이상치로 인한 공분산 행렬의 왜곡 문제 해소에 기여하였지만, 여전히 선형 보간을 통한 오버샘플링의 무분별한 데이터 생성으로 인한 문제를 해결하지 못하였다. 다수 클래스와 소수 클래스가 인접한 영역 또는 영역을 침범한 데이터에 대해서 선형 보간을 통한 오버샘플링이 수행되면 소수 클래스의 인공 데이터가 다수 클래스의 영역과 겹치는 클래스 중첩 문제를 일으키게 된다. 이는 분류 모델의 정밀도(Precision)을 저하시키는 원인이 된다 (Zhang et al., 2023).

따라서 본 연구에서는 데이터의 분포 구조 반영과 이상치를 고려한 공분산 행렬 추정을 위해 선행 연구에서 제안한 방법을 활용하되, 다음과 같은 과정을 통해 기존의 선행 연구와의 차별성을 확보하고자 한다.

첫째, 다수 클래스 영역 침범으로 인한 클래스 중첩 문제를 완화하기 위해 구성된 최근접 이웃 고유 선분과 데이터 분포의 중심점 간의 유사성을 바탕으로 각 선분별 인공 데이터 생성량을 조절한다. 둘째, 유사성이 낮은 선분은 인공 데이터를 적게 생성하여 영역 침범을 최소화하고 분류 경계면의 중첩 문제를 완화한다. 셋째, 유사성이 높은 코어 영역의 선분에는 인공 데이터를 많이 생성하여 소수 클래스의 분포 특성을 보존한다.

3. 이론적 배경

3.1. 데이터 불균형 문제

실제 산업 현장에서 수집되는 데이터는 특정 클래스에 속한 인스턴스의 수가 다른 클래스에 비해 압도적으로 많은 데이터 불균형(Data Imbalance) 문제를 띠는 경우가 많다. 대표적인 사례로 제조 공정의 품질 데이터를 들 수 있는데, 생산된 제품의 대부분은 양품이지만 극히 일부 생산되는 불량품은 양품 대비 개체 수가 현저히 적은 불균형한 구조를 가진다.

이러한 데이터 불균형 상태에서 머신러닝 모델을 학습시킬 경우, 알고리즘은 전체 손실을 최소화하기 위해 다수 클래스를 정확하게 맞추는 방향으로 편향되는 과적합 문제가 발생하기 쉽다. 이 과정에서 나타나는 대표적인 한계점은 '정확도의 역설(Accuracy Paradox)'이다.

정확도(Accuracy)는 전체 예측 시도 중 정답을 맞춘 비율을 의미한다. 데이터의 90%가 양품 데이터일 때 전체 데이터를 양품으로 분류하기만 하여도 90%라는 높은 정확도가 산출된다. 그러나 이러한 모델은 소수 클래스를 분류하는 상황에서 소수 클래스를 하나도 식별해내지 못했음에도 불구하고 수치상으로는 매우 높은 정확도를 나타내게 되어 수치상으로는 매우 우수한 성능을 가진 것처럼 착각하게 만들고 실질적인 분류 모델로서의 가치를 판단하기 어렵게 만든다.

정확도 중심 평가의 한계를 극복하기 위해 예측 결과와 실제 값을 교차한 표 형태로 나타내는 혼동 행렬(Confusion Matrix)을 활용하여 모델의 성능을 평가하기도 한다. [표 3-1]과 같은 혼동 행렬은 모델의 예측 결과를 네 가지 범주로 세분화하여 모델의 성능을 분석할 수 있게 한다.

[표 3-1] 혼동 행렬(Confusion Matrix)

		예측값	
		음성, 0 (Negative)	양성, 1 (Positive)
실제값	음성, 0 (Negative)	True Negative; TN	False Positive; FP (1종 오류)
	양성, 1 (Positive)	False Negative; FN (2종 오류)	True Positive; TP

[표 3-1]의 수치에 대해 설명하면 True Positive(TP)는 실제 양성을 모델이 양성으로 정확하게 예측한 경우이며 True Negative(TN)은 실제 음성을 모델이 음성으로 정확하게 예측한 경우이며 False Positive(FP)는 실제 음성을 모델이 양성으로 잘못 예측한 1종 오류이며 False Negative(FN)는 실제 양성을 모델이 음성으로 잘못 예측한 2종 오류이다.

이 네 가지 기본 요소를 바탕으로 도출되는 주요 성능 지표와 산출 식은 다음과 같다.

1. 정밀도 (Precision): 모델이 양성으로 예측한 대상 중 실제 양성의 비율이다.

$$Precision = \frac{TP}{TP + FP}$$

2. 재현율 (Recall): 실제 양성 중 모델이 양성으로 올바르게 찾아낸 비율이다.

$$Recall = \frac{TP}{TP + FN}$$

3. F1-Score: 정밀도와 재현율의 조화평균으로, 두 지표가 균형을 이루는 정도를 나타낸다.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.2. 데이터 증강 기법

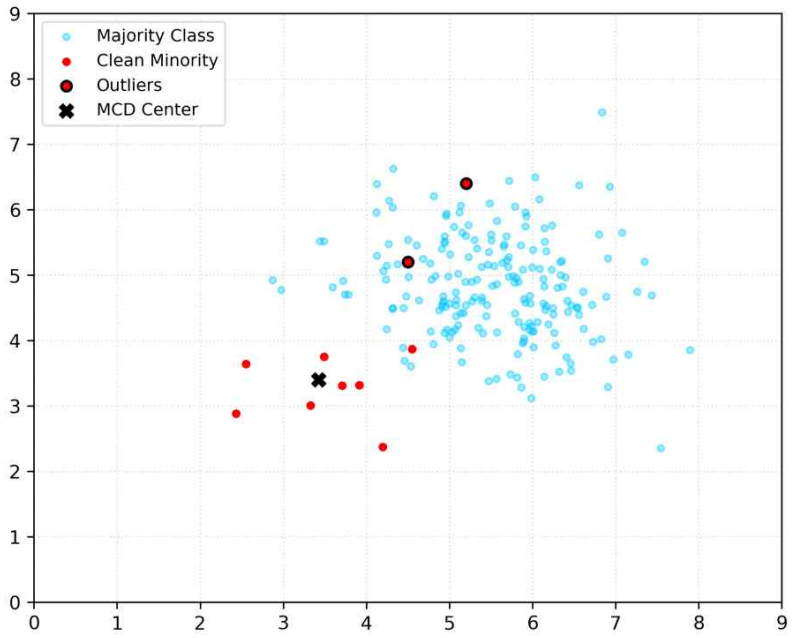
머신러닝 학습에서 데이터 불균형 문제를 해결하기 위한 대표적인 방법으로 데이터 수준 접근법이 있다 (Hulse et al., 2007). 데이터 수준 접근법은 알고리즘 학습 이전에 데이터셋의 다수 클래스와 소수 클래스의 비율을 인위적으로 조정하여 불균형을 해소하는 방법이다. 이는 소수 클래스의 데이터를 복제하거나 새롭게 생성하여 다수 클래스의 비율과 맞추는 오버샘플링(Over Sampling) 기법과 다수 클래스의 데이터를 일부만 남기고 줄여 소수 클래스와 비율을 맞추는 언더샘플링(Under Sampling) 기법, 오버샘플링과 언더샘플링 기법을 혼합하여 사용하는 혼합 샘플링 기법으로 나뉜다.

오버샘플링 기법은 데이터의 추가 생성하는 방식으로 언더샘플링 기법 대비 모델 학습에 걸리는 시간이 기존 보다 증가할 수 있으나 데이터의 중요한 정보 손실을 피할 수 있다는 장점이 있으며 대표적인 오버샘플링 기법으로 Chawla et al.(2002)가 제안한 SMOTE(Synethic Minority Over-Sampling Technique)가 있다 (박정렬, 2021). 언더샘플링 기법은 오버샘플링 기법 대비 계산 효율성은 높지만, 다수 클래스가 가지고 있는 정보를 손실할 수 있다는 리스크가 존재한다. 대표적인 언더샘플링 기법으로 Tomek-Links가 있다 (김한용, 이우주, 2017).

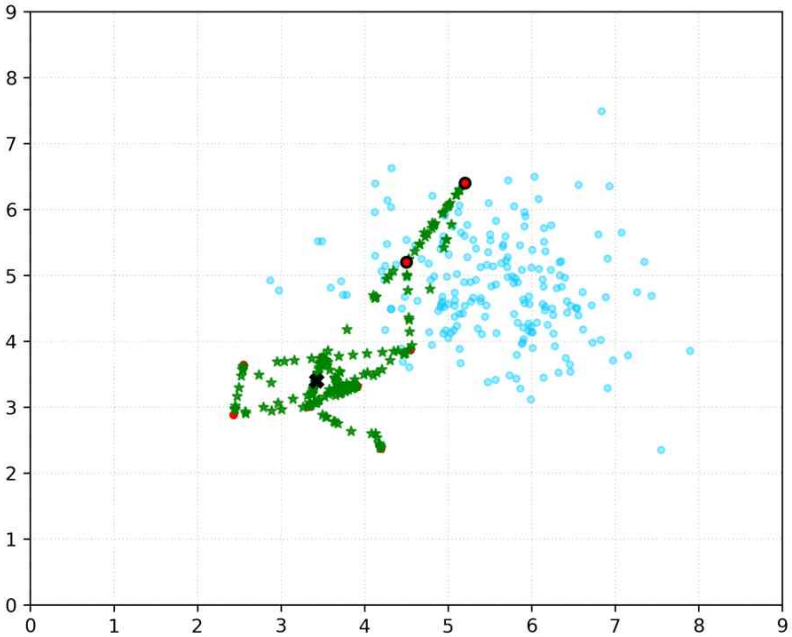
SMOTE는 불균형 데이터 문제 해결을 위해 여러 분야에서 사용되는 대표적인 오버샘플링 기법이다. 해당 기법은 무작위로 인공 데이터를 생성하는 단순 복제 방식의 한계를 극복하기 위해 부트스트래핑(Bootstrapping)과 K-최근접 이웃(K-Nearest Neighbor, KNN) 알고리즘을 결합하여 가상의 소수 클래스 데이터를 생성한다. SMOTE의 생성 메커니즘은 다음과 같이 수행된다. 우선 소수 클래스에 속하는 임의의 데이터 x_i를 선택한 후, 해당 데이터와 동일한 클래스 내에서 거리가 가장 가까운 k개의 이웃 데이터를 탐색한다. 이후 선택된 데이터와 이웃 데이터 사이의 직선 거리를 선형으로 연결하고, 식 (1)과 같이 그 사이의 임의의 지점에 새로운 가상 샘플(x_{new})을 생성한다(Chawla et al., 2002).

$$x_{new} = x_i + rand(0,1) \times (x_j - x_i) \quad (1)$$

이러한 방식은 기존 데이터를 단순히 복제하는 오버샘플링에 비해 소수 클래스의 결정 경계(Decision Boundary)를 외곽으로 확장시켜 모델의 과적합을 방지한다는 장점이 있다.



[그림 3-1] SMOTE 수행 전 원본 데이터 분포



[그림 3-2] SMOTE 수행 후 데이터 분포
(★은 생성된 인공 데이터)

3.3. 마할라노비스 거리(Mahalanobis Distance)

전통적으로 데이터와 데이터 사이의 거리를 표현할 때 사용되는 유클리드 거리(Euclidean Distance)는 모든 변수가 서로 독립적이며 동일한 분산을 가진다는 가정을 전제로 한다. 그러나 실제 다변량 데이터 세트에서는 변수 간의 강한 상관관계가 존재하거나, 축별 스케일 및 산포도가 상이한 경우가 빈번하게 발생한다. 이러한 데이터 분포의 특성을 무시한 채 유클리드 거리를 적용할 경우, 특정 변수의 스케일에 의해 거리 연산 결과가 왜곡되거나 데이터 고유의 기하학적 위상구조를 훼손할 수 있는 한계점이 존재한다.

마할라노비스 거리(Mahalanobis Distance)는 변수 간의 다중공선성과 스케일 차이로 인해 발생하는 공간의 왜곡을 보정하고 다변량 데이터 본연의 통계적 확률 분포를 반영하기 위해 제안되었다. 이 거리 척도는 각 변수의 측정 단위를 자동으로 표준화하는 스케일 불변성의 특징을 가지며, 변수 간의 중복된 정보를 제거하여 순수한 기하학적 거리를 산출하는 장점이 있다. 마할라노비스 거리는 다음과 같이 정의된다 (Mahalanobis, 1936).

p차원 공간에서 평균 벡터가 $\mu=(\mu_1,\mu_2,\cdots,\mu_p)^T$ 이고, 공분산 행렬이 $\Sigma\in\mathbb{R}^{p\times p}$인 데이터 집합이 주어졌을 때, 임의의 관측치 벡터 $x=(x_1,x_2,\cdots,x_p)^T$와 중심점 μ 사이의 마할라노비스 거리 $D_M(x)$는 이하의 식 (2)과 같이 정의된다.

$$D_M(x)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2}$$

또한 동일한 확률 공간상에 존재하는 두 개별 관측치 벡터x와 y 사이의 상대적 거리는 식 (2)와 같이 정의된다.

$$D_M(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)} \tag{3}$$

여기서 Σ^{-1}은 공분산 행렬의 역행렬이며, 공간의 왜곡을 보정한다.

4. 연구 방법론

본 연구에서는 정지은과 최용석(2024)의 연구 절차를 참고하여, Rousseeuw와 Van Driessen(1999)이 제안한 Fast-MCD 알고리즘을 기반으로 소수 클래스의 평균벡터와 공분산을 추정한다. 추정된 평균벡터와 공분산을 이용하여 마할라노비스 거리 공간을 구성하고, 그 거리를 바탕으로 합성 표본 생성 영역을 결정한다. Fast-MCD 알고리즘에 대한 설명은 Algorithm 1에 제시하였다(정지은&최용석, 2024).

Algorithm 1 : Fast-MCD

[Step 1] 데이터 행렬 X에서 h개의 관측치들을 샘플링하여 부분행렬H_1을 구성하고, 그 부분집단의 평균벡터 $\overline{\mathbf{x}}_1$과 공분산 행렬 S_1을 계산한다.

[Step 2] $\overline{\mathbf{x}}_1$과 S_1을 사용하여 전체 데이터 행렬 X에 있는 각 관측치에 대해 마할라노비스 거리 $r(\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_1)$을 계산한다.

$$r(\mathbf{x}_i, \overline{\mathbf{x}}_1) = \left\| \mathbf{x}_i - \overline{\mathbf{x}}_1 \right\|_{S_1^{-1}}, \ i = 1, \ldots, n.$$

[Step 3] Step 2에서 계산한 거리들 중 가장 작은 값들을 갖는 h개의 관측치로 부분행렬 H_2를 구성하고, 이에 대한 평균벡터 $\overline{\mathbf{x}}_2$와 공분산 S_2를 계산한다. (이 경우 $|S_1| \geq |S_2|$가 보장된다.)

[Step 4] $|S_l| = 0$ 또는 $|S_l| = |S_{l+1}|$이 될 때까지 Step 2와 Step 3를 반복한다.

[Step 5] 반복을 통해 얻은 부분행렬 H_l의 평균 $\overline{\mathbf{x}}_l$과 공분산 S_l을 사용하여 전체 데이터 행렬 X의 각 관측치에 대해 최종 마할라노비스 거리 $r(\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_l)$을 계산한다.

$$r(\mathbf{x}_i, \overline{\mathbf{x}}_l) = \left\| \mathbf{x}_i - \overline{\mathbf{x}}_l \right\|_{S_l^{-1}}, \ i = 1, \ldots, n, l = 1, \ldots, n-1.$$

위 과정을 통해 구해진 평균벡터$\overline{\mathbf{x}}_{mcd}$와 공분산 행렬 S_{mcd}를 사용하여 계산된 마할라노비스 거리는 다음과 같다.

$$r(\mathbf{x}_i, \overline{\mathbf{x}}_{mcd}) = \left\| \mathbf{x}_i - \overline{\mathbf{x}}_{mcd} \right\|_{S_{mcd}^{-1}} = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}}_{mcd})^T S_{mcd}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}_{mcd})}.$$

출처: 정지은&최용석(2024), SMOTE by Mahalanobis distance using MCD in imbalanced data, The Korean Journal of Applied Statistics, 37(4), 455-465.

본 연구는 이 거리 정보를 활용하여 소수 클래스 샘플 사이의 고유 선분을 구성하고, 각 선분의 중점과 클래스 분포 중심 간 마할라노비스 거리의 역수에 따라 합성 표본 생성량을 확률적으로 배분하는 방식으로 오버샘플링을 수행한다. 고유 선분의 중점과 클래스 분포 중심 간의 마할라노비스 거리의 역수를 가중치로 활용하는 이유는 소수 클래스 고유의 분포 구조를 보존하고, 클래스의 경계에서 중첩이 발생하는 것을 최소화하기 위함이다.

변수 간의 공분산을 반영하는 마할라노비스 공간 상에서 중심점과 인접한 고유 선분은 소수 클래스의 분포 특성을 반영하는 영역이며 (Mahalanobis, 1936; Hyndman, 1996), 분포의 중심점과 거리가 먼 고유 선분은 실제 데이터가 관측될 확률이 극히 희소한 영역일 뿐만 아니라, 다수 클래스 진영과 극도로 인접하여 무분별한 선형 보간 시 데이터 오버랩을 유발할 수 있는 영역이다 (Zhang et al., 2023). 본 연구에서 제시하는 알고리즘에 대한 설명은 Algorithm 2에 제시하였다.

Algorithm 2 : 본 연구에서 제안하는 방법론

[입력] 전체 데이터 (X, y), 생성할 합성 표본 수 $N_{syn}(N_{maj}-N_{min})$, 이웃수 K, Fast-MCD의 support fraction, 정규화 항 λ, 안정화 상수 ϵ

[Step 1] 클래스 분리 및 목표 생성 개수 결정
- 전체 데이터의 타겟 y로부터 X_{min}와 X_{maj}를 분리한다.
- 생성할 합성표본 수 $N_{syn}(N_{maj}-N_{min})$이 만족되면 종료한다.

[Step 2] Fast-MCD 알고리즘 수행
- 소수 클래스 X_{min}에 대해 Fast-MCD를 적용하여 평균벡터 $\overline{\mathbf{x}}_{mcd}$와 공분산 S_{mcd}를 추정한다.
- 수치 안정화를 위해 $S_{mcd} \leftarrow S_{mcd} + \lambda I$를 적용하고, 역행렬 $V = S_{mcd}^{-1}$을 계산한다.

[Step 3] 고유 선분 구성
- X_{min} 내부에서 $V(=S_{mcd}^{-1})$기반 마할라노비스 거리로 각 샘플의 K 최근접 이웃을 찾는다.
- (i, j)쌍을 정렬된 튜플로 합쳐 중복을 제거한 고유 선분 집합 $U = (x_{a_n}, x_{b_n})_{n=1}^{|Uvert}$을 구성한다.

[Step 4] 선분 중심과 MCD 중심 간 거리 계산
- 각 선분 n에 대해 중점 $MP_n = \dfrac{(x_{a_n} + x_{b_n})}{2}$을 계산한다.
- 각 중점과 MCD 중심 간의 마할라노비스 거리 MD_n을 계산한다.

$$MD_n = \sqrt{(\mathrm{MP}n - \bar{x}_{mcd})^T V (MP_n - \bar{x}_{mcd})}.$$

[Step 5] 가중치 및 확률 산정

- 각 선분에 대해 가중치 $w_n = \frac{1}{MD_n + \epsilon}$로 정의한다.
- 각 선분에 대해 확률 $P_n = \frac{w_n}{\Sigma_m w_m}$로 정규화한다.

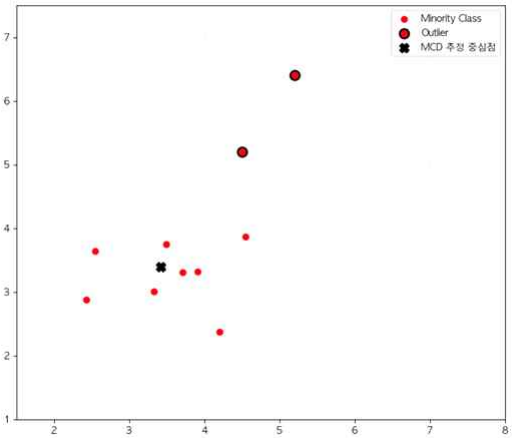
[Step 6] 생성표본 수 배분 및 생성

- N_{syn}을 다항분포 $\mathrm{Multinomial}(N_{syn}; p_1, p_2, \cdots, p_{|U|})$로 분배하여 각 선분에 할당된 생성 수 C_n을 결정한다.
- 각 선분 (x_a, x_b)에서 C_n개를 생성한다. 생성 방식: $u \sim \mathrm{Uniform}(0,1)$, $x_{syn} = x_a + u(x_b - x_a)$

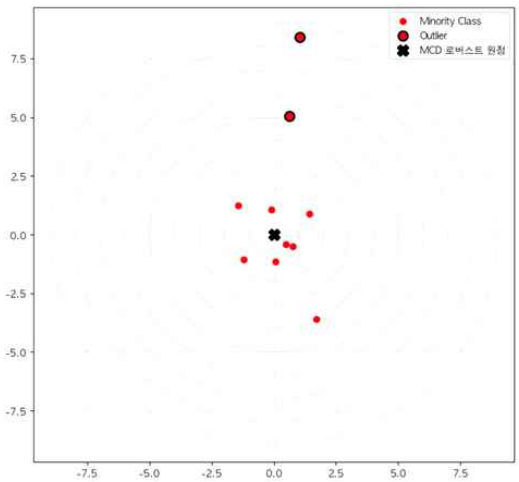
[Step 7] 반환

- 생성된 합성표본들을 원본 데이터에 결합하고 소수 클래스의 레이블을 부여한다.

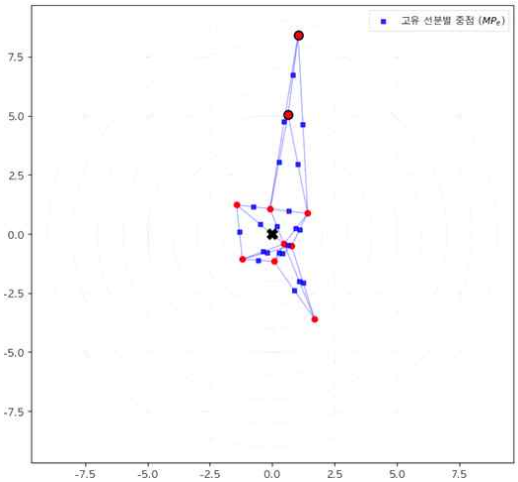
소수 클래스의 유클리드 거리 공간 상 분포가 [그림 4-1]와 같을 때, 마할라노비스 거리 공간 상의 분포는 [그림 4-2]와 같다. 이 때 마할라노비스 거리를 기준으로 각 소수 샘플의 K개의 최근접 이웃을 찾아 이들 간을 잇는 고유 선분을 구성하고, 각 선분의 중점 MP_n을 파란색 점으로 표시한 그림은 [그림 4-3]와 같다. [그림 4-4]은 각 중점 MP_n과 MCD로 추정된 분포 중심점 $\bar{x}_{mcd}$사이의 마할라노비스 거리 $MD_n = \sqrt{(\mathrm{MP}n - \bar{x}_{mcd})^T V (MP_n - \bar{x}_{mcd})}$ 를 나타낸다. 마지막으로 [그림 4-5]는 MD_n의 역수에 비례하는 가중치를 이용하여 다항분포로 각 고유 선분에 할당할 합성 표본의 분포를 나타낸다. 최종적으로 오버샘플링이 완료된 후 전체적인 데이터 분포 구조는 [그림 4-6]과 같다.



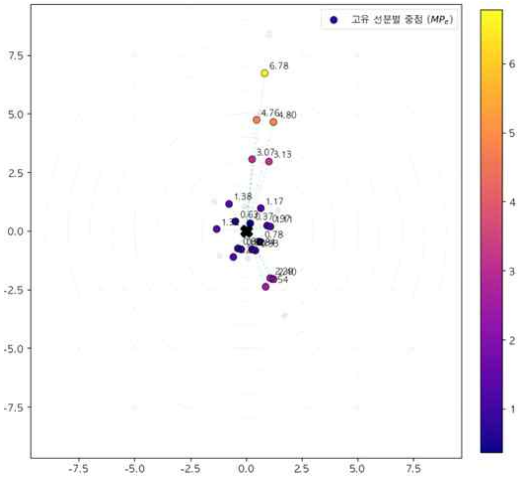
[그림 4-1] 유클리드 거리 공간 내 소수 클래스의 분포



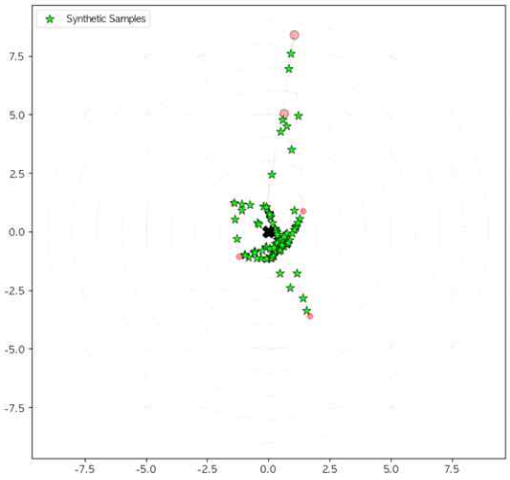
[그림 4-2] 추정된 MCD 마할라노비스 거리 공간 내 소수 클래스의 분포



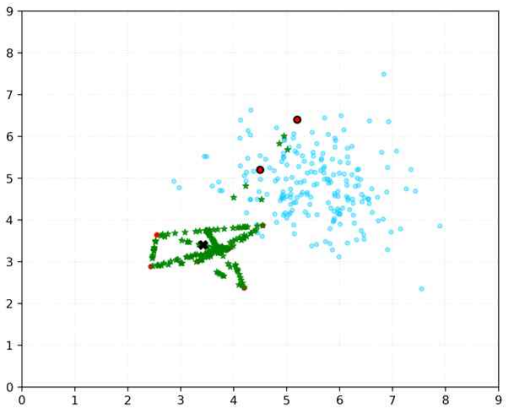
[그림 4-3] 마할라노비스 K-NN 기반 고유 선분과 중점



[그림 4-4] 고유 선분 중점과 MCD 중심 간의 마할라노비스 거리



[그림 4-5] 생성된 고유 선분 별 인공 데이터 분포



[그림 4-6] 오버샘플링 수행 후 유클리드 공간 내 데이터 분포

5. 연구 결과

5.1. 실험 설계

본 연구에서 제안하는 방법론의 성능을 평가하기 위해 'KEEL(Knowledge Extraction based on Evolutionary Learning)' 저장소에서 제공하는 6개의 벤치마크 데이터 세트를 활용하였다 (Alcalá-Fdez et al., 2009). 실험에 사용된 각 데이터 세트에 대한 설명은 [표 5-1]와 같다. 수집된 모든 데이터 세트의 타겟 변수 y는 클래스를 나타내며 두 개의 클래스로 분류되는 이진 분류 데이터이며, 불균형 비율(Imbalance Ratio, IR)이 최소 2.78에서 최대 9.12에 이르는 극단적 불균형 도메인을 포함한다.

[표 5-1] KEEL 벤치마크 데이터 세트

데이터 세트	속성 수	소수 클래스 인스턴스	다수클래스 인스턴스	총 인스턴스	IR
page-blocks0	10	559	4913	5472	8.79
ecoli-0-6-7_vs_3-5	7	22	200	222	9.09
ecoli-0-4-6_vs_5	6	20	183	203	9.15
haberman	3	81	225	306	2.78
ecoli-0-3-4-7_vs_5-6	7	25	232	257	9.28
yeast-0-3-5-9_vs_7-8	8	50	456	506	9.12

실험 수행을 위해 각 데이터 세트는 5-Fold 교차 검증(Stratified K-Fold Cross Validation)을 통해 Train 데이터와 Test 데이터로 분할된다. 분할된 Train 데이터의 소수 클래스는 샘플링을 적용하지 않은 원본 상태(Original), 전통적인 선형 보간 기법인 SMOTE, 정지은과 최용석(2024)이 제안한 MCD-SMOTE, 본 연구에서 제안하는 방법론을 통해 오버샘플링을 수행한다.

이후 샘플링이 완료된 각 Train 데이터를 활용하여 SVM-RBF 모델 학습을 수행하고, Test 데이터를 활용하여 학습된 모델의 성능을 평가한다. 모델에 대한 평가는 모델의 정밀도(Precision)와 재현율(Recall)을 모두 고려해 성능을 객관적으로 평가하기 위해 F1-Score를 활용하여 평가한다.

5.2. 실험 결과

Train 데이터를 5-Fold 교차 검증으로 분할한 후, 각 폴드의 Train 데이터에 오버샘플링 기법을 각각 적용하였다. 이후 샘플링된 데이터를 SVM-RBF 모델에 학습시키고 Test 데이터로 평가하는 과정을 5회 반복하여 모델을 교차 검증 하였다. 각 실험 조건에 따라 도출된 5-Fold F1-Score의 평균과 분산은 [표 5-2]에 제시되어 있으며, Precision과 Recall은 [표 5-3]과 [표 5-4]에 각각 제시되어 있다.

[표 5-2] 벤치마크 데이터 세트의 각 기법 별 F1-Score 평균 및 분산

Dataset	Original	SMOTE	MCD-SMOTE	Proposed Method
page-blocks0	0.1129 (0.0013)	0.2728 (0.0001)	0.2696 (0.0002)	**0.3879** (0.0008)
ecoli-0-6-7_vs_3-5	**0.7548** (0.0037)	0.6877 (0.0333)	0.6842 (0.0103)	0.7175 (0.0092)
ecoli-0-4-6_vs_5	0.8262 (0.0127)	0.8192 (0.0184)	0.8414 (0.0234)	**0.8643** (0.0063)
haberman	0.0211 (0.0018)	0.3791 (0.0029)	0.3930 (0.0022)	**0.4243** (0.0166)
ecoli-0-3-4-7_vs_5-6	0.8018 (0.0194)	**0.8240** (0.0132)	**0.8240** (0.0132)	**0.8240** (0.0132)
yeast-0-3-5-9_vs_7-8	0.2953 (0.0342)	0.3545 (0.0056)	0.3631 (0.0041)	**0.3856** (0.0036)

[표 5-3] 벤치마크 데이터 세트의 각 기법 별 Precision 평균 및 분산

Dataset	Original	SMOTE	MCD-SMOTE	Proposed Method
page-blocks0	**0.8381** (0.0140)	0.1598 (0.0001)	0.1577 (0.0001)	0.2668 (0.0006)
ecoli-0-6-7_vs_3-5	**0.9500** (0.0100)	0.6633 (0.0434)	0.6310 (0.0075)	0.7989 (0.0418)
ecoli-0-4-6_vs_5	**0.9500** (0.0100)	0.8100 (0.0344)	0.8500 (0.0400)	**0.9500** (0.0100)
haberman	0.0667 (0.0178)	0.4688 (0.0189)	**0.4765** (0.0145)	0.4393 (0.0218)
ecoli-0-3-4-7_vs_5-6	**0.8867** (0.0247)	0.8295 (0.0248)	0.8295 (0.0248)	0.8295 (0.0248)
yeast-0-3-5-9_vs_7-8	**0.7333** (0.1511)	0.2466 (0.0029)	0.2482 (0.0019)	0.2784 (0.0019)

[표 5-4] 벤치마크 데이터 세트의 각 기법 별 Recall 평균 및 분산

Dataset	Original	SMOTE	MCD-SMOTE	Proposed Method
page-blocks0	0.0608 (0.0004)	**0.9338** (0.0004)	0.9320 (0.0005)	0.7138 (0.0012)
ecoli-0-6-7_vs_3-5	0.6400 (0.0094)	0.7300 (0.0296)	**0.7800** (0.0416)	0.7400 (0.0464)
ecoli-0-4-6_vs_5	0.7500 (0.0250)	**0.8500** (0.0150)	**0.8500** (0.0150)	0.8000 (0.0100)
haberman	0.0125 (0.0006)	0.3338 (0.0012)	0.3456 (0.0008)	**0.4412** (0.0191)
ecoli-0-3-4-7_vs_5-6	0.7600 (0.0384)	**0.8400** (0.0224)	**0.8400** (0.0224)	**0.8400** (0.0224)
yeast-0-3-5-9_vs_7-8	0.2000 (0.0200)	0.6400 (0.0224)	**0.6800** (0.0176)	0.6400 (0.0224)

정량적 비교 분석 결과, 본 연구에서 제안한 방법론을 적용한 모델의 F1-Score가 page-blocks0, ecoli-0-4-6_vs_5, haberman, yeast-0-3-5-9_vs_7-8 등 다수의 극단적 불균형 지형에서 대조군 대비 높은 성능을 기록하였다. 특히 page-blocks0 데이터 세트의 경우, 원본 모델 및 기존 오버샘플링 모델들 대비 높은 F1-Score를 기록하였다.

F1-Score는 정밀도(Precision)와 재현율(Recall)의 조화평균이라는 점을 고려하여 정밀도와 재현율을 비교하였다. 정밀도는 제안한 방법론이 page-blocks, ecoli-0-6-7_vs_3-5, ecoli-0-4-6_vs_5, yeast-0-3-5-9_vs_7-8 데이터 세트에서 대조군인 SMOTE, MCD-SMOTE 대비 높은 정밀도를 보였다. 이는 제안한 방법론이 대조군 방법론 대비 클래스 중첩 문제를 완화하였음을 의미한다. 한편, 샘플링을 적용하지 않은 원본(Original) 데이터에서 높은 정밀도가 나오는 것은 분류기가 다수 클래스로 편향되어 소수 클래스 예측 분모 자체를 극소화하여 발생하는 불균형 데이터 고유의 문제점인 정확도의 역설에 의한 것이다.

반면에 재현율은 제안한 방법론이 비교하는 방법론 대비 높은 향상을 보이지 못하였다. 이는 마할라노비스 거리의 역수 가중치에 따른 고유선분 별 인공 데이터 생성량 조절 기법의 한계점으로, 확률 밀도가 높은 소수 클래스의 중심부 영역에 인공 데이터 생성량이 집중되어 클래스 경계 인근 데이터 생성량이 부족하여 분류 모델의 클래스 경계 영역 학습이 부족하였다는 점을 확인할 수 있다. 향후 연구에서는 재현율의 향상을 위해 생성량을 조절하는 가중치 함수를 고도화할 필요가 있다.

5.3. 통계적 유의성 검정

본 연구에서는 제안한 방법론의 높은 F1-Score 순위가 우연에 의한 것인지 검정하기 위해 6종의 F1-Score 평균치를 기반으로 프리드만 검정(Friedman Test)과 사후 검정으로 윌콕슨 부호순위 검정(Wilcoxon Signed-Rank Test)을 수행하였고, 프리드만 검정의 절차는 다음과 같다.

단계 1: 가설 설정

H_0 : 비교 대상인 알고리즘 간의 성능 순위 격차는 존재하지 않는다.
H_1 : 비교 대상 알고리즘 간의 성능 순위는 통계적으로 유의미한 차이가 존재한다.

단계 2: 검정통계량 및 유의확률 계산

$\chi^2 = \chi_0^2$(검정통계량)일 때, $P-value$(유의확률) $= p^*$

단계 3: 가설 검정

$P-value < \alpha$(유의수준) 일 때, H_0 기각

실험을 통해 얻은 F1-Score 평균을 활용하여 프리드만 검정을 수행한 결과 검정통계량 $\chi_0^2 = 8.3571$와 이 때의 유의 확률 $p^* = 0.046875$를 얻었다. 유의확률 $p^* = 0.046875$가 유의수준 $\alpha = 0.05$보다 작기 때문에 귀무가설 H_0을 기각한다. 즉, 비교 알고리즘 전체 간의 등수 차이에 통계적으로 유의미한 차이가 존재한다고 볼 수 있으므로, 제안 방법론을 적용한 모델의 성능이 유의하게 우월하다는 것을 확인하기 위해 사후 검정으로 윌콕슨 부호순위 검정을 수행하였다. 윌콕슨 부호순위 검정 절차는 다음과 같다.

단계 1: 가설 설정

H_0 : 대조군 알고리즘과 제안 방법론 간의 성능 차이가 없거나 대조군의 성능이 더 좋다.
H_1 : 대조군 알고리즘보다 본 연구에서 제안하는 방법론의 성능이 더 좋다.

단계 2: 검정통계량 및 유의확률 계산

$W = w_0$(검정통계량) $= \sum_{D_i > 0} \mathrm{Rank}(|D_i|)$ 일 때, $P-value$(유의확률) $= p^*$

단계 3: 가설 검정

$P-value < \alpha$(유의수준) 일 때, H_0 기각

[표 5-5] 윌콕슨 부호 검정 결과

	Original	SMOTE	MCD-SMOTE
윌콕슨 통계량 (w_0)	$w_0 = 19.0$	$w_0 = 15.0$	$w_0 = 15.0$
사후 단측 $P-$value	$p^* = 0.046875$	$p^* = 0.031250$	$p^* = 0.031250$
귀무 가설 기각 여부 ($\alpha = 0.05$)	$0.046775 < 0.05$ H_0 기각	$0.031250 < 0.05$ H_0 기각	$0.031250 < 0.05$ H_0 기각

[표 5-5]은 Original, SMOTE, MCD-SMOTE를 대상으로 본 연구의 제안 방법론과 각각 사후 윌콕슨 부호순위 검정의 결과를 나타낸다.

모든 개별 대조군은 사후 단측 유의확률(p^*)이 유의수준 $\alpha = 0.05$보다 모두 작으므로 모든 대조군에 대해 귀무가설을 기각한다. 즉, 제안 방법론을 적용한 모델이 기존의 방법론을 적용한 모델 대비 평균 F1-Score의 수치가 유의미하게 높다는 것을 통계적으로 검증하였다.

6. 결론 및 향후 과제

본 연구는 데이터 불균형 문제를 해결하기 위한 새로운 오버샘플링 기법을 제안하였다. 기존의 SMOTE는 유클리드 거리 기반의 선형 보간을 수행함으로써 데이터의 공분산 구조를 반영하지 못하며, 다수 클래스 영역을 침범하는 클래스 중첩 문제를 유발할 수 있다는 한계를 가진다. 또한 마할라노비스 거리를 활용한 MDO와 MCD-SMOTE는 데이터의 분포 구조 반영 및 이상치 문제를 일부 개선하였으나, 여전히 모든 선분에 대해 균일한 선형 보간을 수행함으로써 클래스 중첩 문제를 해결하지 못하였다.

이에 본 연구에서는 Fast-MCD를 활용하여 소수 클래스의 평균 벡터와 공분산 행렬을 추정하고, 이를 기반으로 구성된 마할라노비스 거리 공간에서 최근접 이웃 간의 고유 선분을 생성하였다. 이후 각 고유 선분의 중점과 MCD 중심 간의 마할라노비스 거리를 계산하고, 거리의 역수를 가중치로 활용하여 선분별 인공 데이터 생성량을 차등 배분하는 새로운 오버샘플링 방법론을 제안하였다. 이를 통해 소수 클래스의 핵심 영역에는 상대적으로 많은 데이터를 생성하고, 분포의 경계 영역이나 이상치 주변에는 적은 데이터를 생성함으로써 클래스 중첩을 완화하고 분포 구조를 보존하고자 하였다.

KEEL 저장소의 6개 불균형 벤치마크 데이터 세트를 활용한 실험 결과, 제안 방법론은 대부분의 데이터 세트에서 기존 SMOTE 및 MCD-SMOTE보다 높은 F1-Score를 나타냈다. 또한 프리드만 검정과 윌콕슨 부호순위 검정을 수행한 결과, 제안 방법론의 성능 향상이 통계적으로 유의함을 확인하였다. 이를 통해 본 연구에서 제안한 방법론이 클래스 중첩 문제를 완화하며 불균형 데이터 분류 성능 향상에 효과적임을 검증하였다.

그러나 본 연구에도 한계점이 존재한다. 현재 제안 방법은 고유 선분의 중점과 중심 간 거리의 역수를 직접 가중치로 사용하기 때문에 중심부에 생성 확률이 과도하게 집중될 가능성이 존재한다. 이 경우 소수 클래스의 핵심 영역은 더욱 강화되지만, 상대적으로 경계 영역에 대한 학습이 부족해져 재현율(Recall)이 제한될 수 있다. 실제 실험에서도 일부 데이터 세트에서는 기존 방법론 대비 재현

율 향상이 크지 않은 현상이 관찰되었다.

향후 연구에서는 거리의 역수 대신 비선형 가중치 함수 또는 밀도 기반 가중치 함수를 적용하여 생성 확률의 편중 현상을 완화할 필요가 있다. 또한 현재 연구는 이진 분류 문제를 대상으로 수행되었으므로 다중 데이터 불균형 문제에 대한 확장 가능성을 검토할 필요가 있으며, 6종의 제한된 벤치마크 데이터 세트만을 활용하였고 다수의 분류기를 통한 비교 검증이 부족하였기 때문에 향후 연구에서는 다수의 데이터 세트와 다양한 분류기를 활용한 일반화 성능 검증이 필요하다.

참고 문헌

김한용, 이우주. (2017). 불균형적인 이항 자료 분석을 위한 샘플링 알고리즘들: 성능비교 및 주의점. 응용통계연구, 30(5), 681-690.

박정렬. (2021). 불균형 데이터 문제 해결을 통한 이직의도 설명력 향상 방안에 관한 연구: SMOTE 및 생성적 적대 신경망을 중심으로 (박사학위논문, 충북대학교 대학원).

정지은, 최용석. (2024). 불균형 자료에서 MCD를 활용한 마할라노비스 거리에 의한 SMOTE. 응용통계연구, 37(4), 455-465.

Abdi, L., & Hashemi, S. (2016). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering, 28*(1), 238-251.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Herrera, F., ... & Bacardit, J. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307-318.

Bennin, K. E., Toda, K., Kamei, Y., Monden, A., & Ubayashi, N. (2018). MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance problem in software defect prediction. *IEEE Transactions on Software Engineering, 44*(6), 534-550.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Hulse, J. V., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 935-942.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician, 50*(2), 120-126.

Krishna, S., & Sidharth, S. (2022). HR analytics: Employee attrition analysis using random forest. *International Journal of Performability Engineering, 18*(4), 275.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*(1), 49-55.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223.

Vinoodhini, D. (2022, April). Effective classification of IBM HR analytics employee attrition using sampling techniques. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-6). IEEE.

Zhang, R., Lu, S., Yan, B., Yu, P., & Tang, X. (2023). A density-based oversampling approach for class imbalance and data overlap. *Computers & Industrial Engineering, 186*, 109747.